

HOW IT WORKS

Search Engines

Search engines are simple to use and make online browsing effortless, so it's easy to forget how hard they work and how daunting the internet would be without them. Mike James marvels at how they do it

The web would be unusable if we didn't have effective search engines to help us find the information we want. The web currently comprises more than 4,000,000,000,000 pages, most of it disorganised and of dubious quality. The modern search engine helps us find relevant material, and high-quality relevant material at that. This would seem to be a near-impossible task, but it's essential if the internet is to continue expanding. Knowing how search engines work can help you find the pages you're looking for and, if you run your own website, attract more visitors.

ALONG CAME A SPIDER

To return your search results quickly, search engines maintain an index of the web on their servers and run the search on that. Keeping the index up to date means continually exploring the web, looking for changes and new pages. They do this by automating exactly what you do when you surf the web. Software robots or 'bots' surf the web automatically by following links from one page to another. Bots are also known as spiders, crawlers, worms and agents. This gives the impression that the software roams around the web itself. However, it only ever runs on a single machine and simply downloads thousands of web pages.

It is difficult to design an efficient bot that will crawl the web effectively. You have to account for the possibility that pages have links to each other that loop round in a circle. An unintelligent bot could end up cataloguing the same pages over and over again. There is also the issue of where to start

the search. Most search engines allow website owners to submit their URLs through an online form, and the bots use these as new landing points.

INDEX ISSUES

No software is as intelligent as a human browser: the bot cannot recognise images on a web page or read and understand the sense of sentences. After downloading a page, the bot indexes it based simply on the appearance and frequency of the words it contains. But most of these words won't provide much of a clue as to the page's subject.

Smarter automated search engines try to be clever about which words they use in the index to improve the quality of the cataloguing. For example, words that appear in titles are likely to be more important than words in the body text. The search engine Lycos is said to index words in titles, subheadings and links, together with the 100 most frequently used words on the page plus every word in the first 20 lines of text.

In an effort to help search engines, new Hypertext Mark-up Language (HTML) tags called meta tags were introduced that allowed web page creators to categorise pages themselves. Sadly, meta tags can be used dishonestly to suggest that a page should be returned when it's only trying to sell something. Nowadays bots use meta tags only if the page content seems to agree with the tag.

RANK INSIDER

The real breakthrough in search engine design was PageRank. This was invented by Stanford University

students Sergey Brin and Lawrence Page, who went on to create Google. PageRank tries to rate the quality and relevance of a page in the same way a person does. The rating is based on the pages that link to a page: the more pages that link to a given page, the more important it must be. However, PageRank also takes into account the nature of each linking page, noting their individual PageRanks and the number of other links they carry.

You can think of PageRank as a credibility rating. A link spreads one page's credibility to another, but it's shared out, so a page with lots of links spreads its credibility thin. A page's credibility is the sum of all the credibility passed to it through links. So a page can have a high PageRank either because a small number of high-ranking pages link to it or because lots of low-ranking pages link to it.

GOOGLEWHACKING

The sport of Googlewhacking involves typing in two words that cause Google to return exactly one result – that is, a single web page on which both words appear, or a 'pure whack'. If you achieve this, and it is quite hard, you can submit it to www.googlewhack.com. Of course, as soon as the pure whack is listed on another website it ceases to be one because Google indexes the new page and returns more than one result.

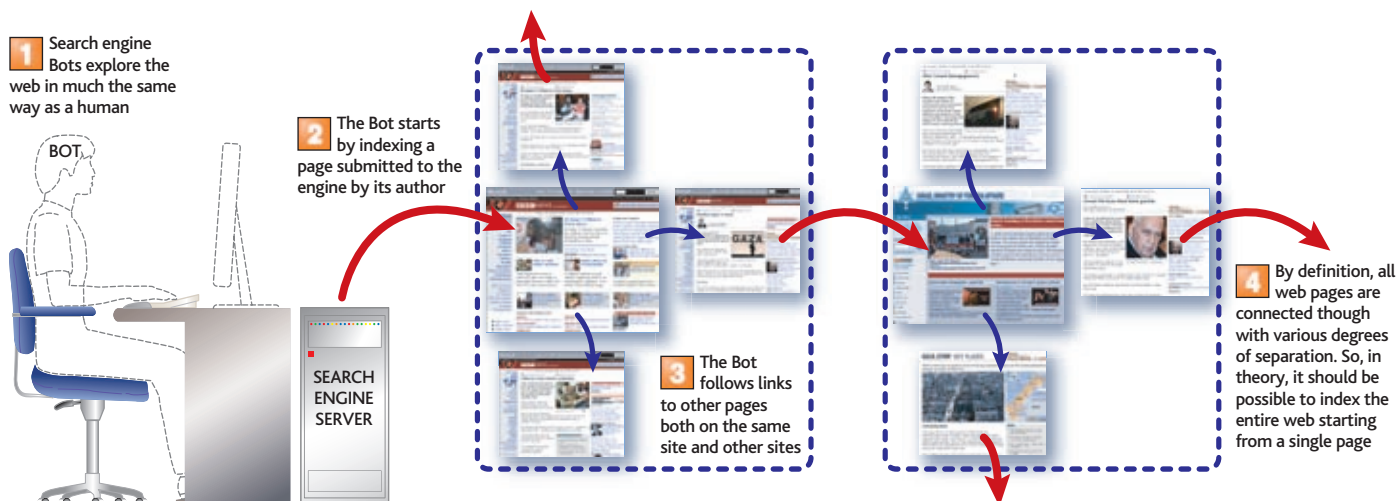
HOW BOTS INDEX THE WEB

1 Search engine
Bots explore the
web in much the same
way as a human

2 The Bot starts
by indexing a
page submitted to the
engine by its author

3 The Bot
follows links
to other pages
both on the same
site and other sites

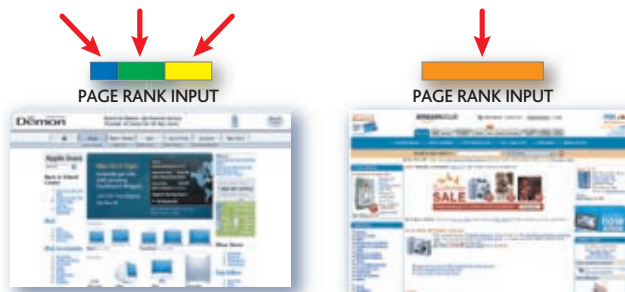
4 By definition, all
web pages are
connected though
with various degrees
of separation. So, in
theory, it should be
possible to index the
entire web starting
from a single page



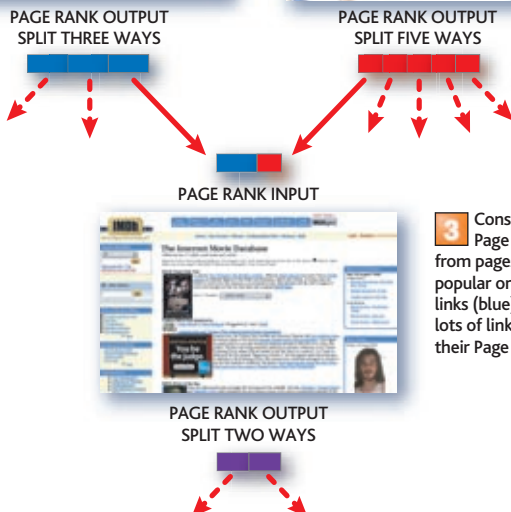


HOW GOOGLE'S PAGE RANKS WORK

1 A page's Google Page Rank is calculated according to what other pages link to it. This page receives Page Rank input from three linking pages



2 A page's Page Rank is passed on to other pages it links to, but only after being divided up according to the number of links



3 Consequently, more Page Rank is passed from pages that are either popular or contain few links (blue). Pages with lots of links (red) spread their Page Rank thinly

Computing a PageRank isn't easy because of the huge effect of any single link being made or broken. It requires a mathematical process called iteration to propagate PageRank changes through the index. You can find out more about Google's origins and PageRank's design at www-db.stanford.edu/~backrub/google.html.

SEARCHING TIMES

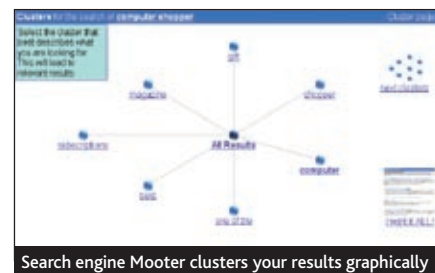
Search engines are at war both with each other and with the people who try to manipulate them to deliver their own pages. A huge amount of time is invested in delivering better, more relevant search results. Yet despite the intense competition, PageRank – or developments based on it – is still the best method we have. It has acquired various tricks to ensure it works better, plus traps and checks that counter attempts to make pages rank highly that don't deserve it. As this is a commercial war, details are hard to come by;

secrecy is the best defence against those who would distort search results.

One big problem facing all search engines is the invisible web, or 'deep web'. As the technology used to create the web gets more sophisticated, more and more web pages are generated dynamically from databases. These are often impossible for bots to find by following links. Current estimates suggest that the deep web holds 500 times more information than the visible web.

A new generation of bots, called wrappers, has been designed to search the deep web. Deep Query Manager, for example, is a search engine that can interact with databases and other search engines to find pages that would otherwise be missed. A search engine that does this is known as a metasearch engine. Deep Query Manager goes much further in searching dynamic databases, though. The latest version of MSN Search also uses the deep web to return information from the

of which holds a 'shard' of the web. When a user's query arrives, it is sent to each shard. The top 1,000 results come back as document ID numbers. The documents are retrieved from Google's document servers, which store a copy of the pages retrieved by the Google's web bots. The page the user sees is assembled by an ad server, which drops in appropriate advertising and sponsored links so that Google makes money and you don't have to pay for a search. All Google's software has been built from scratch on top of Red Hat Linux.



Search engine Mooter clusters your results graphically

Encarta encyclopedia, news services and even an automatic calculator and equation solver.

GENERAL CLUSTER

Another important area of development is the way results are presented. Instead of returning pages in rank order, a new family of search engines employs clustering to present the user with a set of choices to refine the search. North Light and Clusty have pioneered the method, which involves comparing the content of search result pages and giving the user categorised results. More recently, the Mooter search engine added a graphical display that allows users to navigate through pages with similar content. Graphical presentations are even used to display the hidden information in PageRank type relationships. The Kartoo search engine draws a map of the results according to how important they are and shows the sites that refer to them.

SMART MOVE

The use of artificial intelligence may be of greater practical benefit to everyday users. Some new search engines are trying to incorporate an understanding of images and sounds to return web pages with content similar to a sketch or a snatch of music the user provides. Currently, though, AI isn't good enough to help you find what you want with anything approaching human intelligence.

A better idea is to use personal information to tailor a search. Both Ask Jeeves and Amazon's A9 search engine keep a history of your searches and suggest pages that might be particularly relevant to you. A personalised version of Google lets users specify what they find interesting to garner a more appropriate selection from the pages returned. A similar idea uses information from a global positioning system to suggest local shops and resources that might be of interest.

The worry with personalised search is privacy. Soon search engines may be able to examine every web page you look at, keep a list of everything you buy and scan your email for tell-tale signs of your interests. This information could be valuable to anyone who wants to sell to you, which would take search engine advertising into a new realm. **ES**

GOOGLE'S HARDWARE

Google invented PageRank and has implemented an amazing system of hardware to make it all work. Currently Google indexes over 4 million million web pages, each of which is an average of 10K in size. To do this, and to respond to users in a split-second, requires a network based on 30 clusters of machines, with up to 2,000 machines in each cluster. Each cluster can store around a million gigabytes of data. The machines are low-cost servers, and redundancy ensures the system keeps working if a single machine fails. The entire web index is stored on multiple machines, each

FURTHER INFORMATION

Fanciful account of how Google works
www.google.com/technology/pigeonrank.html

General search engine information
www.searchengineshowdown.com

Deep Query Manager
www.brightplanet.com

Mooter
www.mooter.com